

Minimax optimal convex methods for Poisson inverse problems under ℓ_q -ball sparsity

Yuan Li^{*} and Garvesh Raskutti[†]

Abstract

In this paper, we study the minimax rates and provide a convex implementable algorithm for Poisson inverse problems under weak sparsity and physical constraints. In particular we assume the model $y_i \sim \text{Poisson}(Ta_i^\top f^*)$ for $1 \leq i \leq n$ where $T \in \mathbb{R}_+$ is the intensity, and we impose weak sparsity on $f^* \in \mathbb{R}^p$ by assuming f^* lies in an ℓ_q -ball when rotated according to an orthonormal basis $D \in \mathbb{R}^{p \times p}$. In addition, since we are modeling real physical systems we also impose positivity and flux-preserving constraints on the matrix $A = [a_1, a_2, \dots, a_n]^\top$ and the function f^* . We prove minimax lower bounds for this model which scale as $R_q(\frac{\log p}{T})^{1-q/2}$ where it is noticeable that the rate depends on the intensity T and not the sample size n . We also show that a convex ℓ_1 -based regularized least-squares estimator achieves this minimax lower bound up to a $\log n$ factor, provided a suitable restricted eigenvalue condition is satisfied. Finally we prove that provided n is sufficiently large, our restricted eigenvalue condition and physical constraints are satisfied for random bounded ensembles. Our results address a number of open issues from prior work on Poisson inverse problems that focusses on strictly sparse models and does not provide guarantees for convex implementable algorithms.

^{*}Departments of Statistics, University of Wisconsin-Madison, 1300 University Avenue, Madison, WI 53706. The research of Yuan Li is supported in part by NSF Grant DMS-1407028

[†]Departments of Statistics and Computer Science, and Optimization Group at Wisconsin Institute for Discovery, University of Wisconsin-Madison, 1300 University Avenue, Madison, WI 53706. The research of Garvesh Raskutti is supported in part by NSF Grant DMS-1407028

1 Introduction

Large-scale Poisson inverse problems arise in a number of applications where counts are modeled using a Poisson distribution. Examples include imaging (see e.g. Duarte et al. (2008); Studer et al. (2012)), conventional fluorescence microscopy (see e.g. Harmany et al. (2011); Bobin et al. (2007b,a)), network flow analysis (see e.g. Estan and Varghese (2003); Lu et al. (2008); Raginsky et al. (2010a, 2011)), DNA analysis (see e.g. Sansonnet (2013)) and there are many more. In all these problems, a small number of events (e.g. photons hitting a sensor, packets being output, etc.) are observed and these are modeled using a Poisson distribution. In many of these applications, the number of observed events is small relative to the number of model parameters meaning we are in the so-called *high-dimensional* setting.

One standard approach to model the observed counts in the settings above is via a high-dimensional Poisson inverse problem. Specifically, $(y_i)_{i=1}^n$ follows a Poisson distribution and if $y = (y_1, y_2, \dots, y_n)^\top \in \mathbb{R}^n$, we consider the Poisson linear model defined as (see e.g. Raginsky et al. (2010b); Jiang et al. (2014); Willett and Raginsky (2011)):

$$y \sim \text{Poisson}(TAf^*), \quad (1)$$

where $A \in \mathbb{R}_+^{n \times p}$ is a sensing matrix corresponding to the n different projections of our signal of interest $f^* \in \mathbb{R}_+^p$ and $T \in \mathbb{R}_+$ is the total intensity. In particular, (1) is a shorthand expression for the model

$$y_i \sim \text{Poisson}\left(T \sum_{j=1}^p A_{ij} f_j^*\right), \quad i = 1, \dots, n,$$

where the y_i 's are independent. Here our goal is to learn the underlying parameter vector f^* based on the observed counts $(y_i)_{i=1}^n$ where A is known. Furthermore $p \gg n$ since we are in the high-dimensional setting.

Since we are interested in modeling real physical systems corresponding to the applications described above, additional physical constraints are required on f^* and A as in Raginsky et al. (2010b); Jiang et al. (2014); Willett and Raginsky (2011). Since f^* corresponds to the rate at which events occur, $f^* \succeq 0$. Further, we impose the normalization $\|f^*\|_1 = 1$ which is also used in Jiang et al. (2014). In addition A must be composed of

non-negative real numbers with each column summing to at most one. Specifically A must satisfy the following physical constraints:

$$A \succeq 0 \quad (2)$$

$$A^\top \mathbb{1}_{n \times 1} \preceq \mathbb{1}_{p \times 1}. \quad (3)$$

The first constraint (2) is referred to as *positivity* that is $A_{ij} \geq 0 \forall (i, j)$ which ensures that $Af^* \succeq 0$ provided $f^* \succeq 0$ and the second constraint (3) corresponds to a *flux-preserving* constraint which ensures that $\|Af^*\|_1 \leq \|f^*\|_1$. Both constraints make intuitive sense and are natural for the applications described above since counts must be non-negative and the output flux or energy can not exceed the input flux or energy.

Recent work of Jiang et al. (2014) provides minimax optimal rates for Poisson inverse problems described here, when f^* is *strictly sparse* in a basis spanned by the columns of the orthonormal matrix $D \in \mathbb{R}^{p \times p}$, meaning that only a small subset of entries of $D^\top f^*$ are non-zero. Furthermore, Jiang et al. (2014) provides only minimax upper and lower bounds and do not provide theoretical guarantees for implementable convex methods.

In many scenarios (e.g. imaging, network flow analysis), the signal of interest $D^\top f^*$ is *weakly sparse* meaning that low-dimensional structure is imposed on $D^\top f^*$ by requiring that its co-efficients need not be zero, but many co-efficients make a very small contribution to the overall signal. To be precise we write $D = [d_1, \dots, d_p]$ for $d_j \in \mathbb{R}^p$ for all j and assume that $d_1 = p^{-1/2} \mathbb{1}_{p \times 1}$ and denote $\bar{D} = [d_2, \dots, d_p] \in \mathbb{R}^{p \times (p-1)}$. We define $\theta^* = D^\top f^* \in \mathbb{R}^p$, and by construction $\theta_1^* = 1/\sqrt{p}$. To impose weak sparsity, we require the signal $\bar{\theta}^* \triangleq \bar{D}^\top f^* \in \mathbb{R}^{p-1}$ to lie in an ℓ_q -ball, meaning $\|\bar{\theta}^*\|_q^q = \|\bar{D}^\top f^*\|_q^q = \sum_{j=1}^{p-1} |(\bar{D}^\top f^*)_j|^q \leq R_q$ where $0 < q \leq 1$. In this paper, we study the Poisson model (1) under the positivity and flux-preserving constraints and ℓ_q -ball sparsity.

To summarize, we assume f^* belongs to the following set:

$$\mathcal{F}_{p,q,D} = \{f \in \mathbb{R}_+^p : \|f\|_1 = 1, \|\bar{D}^\top f\|_q^q \leq R_q\}.$$

Note that our ℓ_q -ball assumption ensures that many of the co-efficients are small and the convention that is often used is $q = 0$ corresponds to the strictly sparse case studied in Jiang et al. (2014). In this paper, we study minimax rates for the mean-squared ℓ_2 -error for

the Poisson inverse problem (1) where f^* lies in $\mathcal{F}_{p,q,D}$ where $0 < q \leq 1$. That is we provide (1) a lower bound with high probability on the following quantity:

$$\min_{f(A,y)} \max_{f^* \in \mathcal{F}_{p,q,D}} \|f - f^*\|_2^2,$$

where the minimum is taken over measurable functions of (A, y) ; and (2) we show that a convex ℓ_1 -penalized approach achieves this optimal rate up to a $\log n$ factor.

1.1 Our Contributions

Our paper makes three contributions that are novel theoretical results related to this model:

- Provide a minimax lower bound which scales as $R_q\left(\frac{\log p}{T}\right)^{1-q/2}$.
- Show that our minimax lower bound can be achieved up to a $\log n$ factor by an ℓ_1 -based convex method under a suitable restricted eigenvalue condition.
- We prove that random bounded ensembles satisfy the restricted eigenvalue condition along with the imposed physical constraints.

Our bounds are consistent with the intuition from Jiang et al. (2014) under strictly sparse models, where the intensity T and not the sample size n influences the minimax rate. To further support this intuition we provide a comparison of our result to the linear Gaussian inverse problem studied in Raskutti et al. (2011), and show how the minimax rates match the linear Gaussian rate when we set the noise variance σ^2 in terms of n and T in Section 3.1.

We point out that it is not straightforward to adapt the techniques developed in Jiang et al. (2014) to the ℓ_q -ball case. The techniques we use involve combining techniques for proving minimax rates in the high-dimensional Gaussian linear inverse problems under weak sparsity used in Raskutti et al. (2011) and theoretical results for convex implementable methods developed in Negahban et al. (2012) to the linear Poisson setting. A number of technical challenges arise in analyzing the Poisson inverse problem setting since the noise is now signal-dependent. In particular, to use techniques from Negahban et al. (2012) in the Poisson setting we need to use and develop two-sided concentration bounds for Poisson inverse

problems which build on prior work in Bobkov and Ledoux (1998). We go into greater detail on the technical challenges in Sections 2, 3 and 4.

The remainder of this paper is organized as follows: In Section 2 we provide our main assumptions and theoretical results which includes a minimax lower bound, an upper bound for convex methods and discuss matrices A that satisfy the assumptions leading to minimax rates; in Section 3 we discuss a number of comparisons and implications of our results in particular, comparisons to the linear Gaussian model studied in Raskutti et al. (2011) and a comparison to the strictly sparse case in Jiang et al. (2014). Proofs are provided in Section 4.

2 Assumptions and Main Results

In this section, we present our assumptions and main results, which includes a minimax lower bound, an upper bound for a convex ℓ_1 -based approach that matches the minimax lower bound up to a $\log n$ factor and finally we show that if A is a random matrix with suitably bounded entries, it satisfies the statistical conditions and physical constraints.

2.1 Minimax Lower Bound

We begin by introducing the Assumptions for the minimax lower bound.

Assumption 2.1. There exists constants a_ℓ and a_u such that $a_\ell < a_u$ and a matrix $\tilde{A} \in [a_\ell/\sqrt{n}, a_u/\sqrt{n}]^{n \times p}$

$$A = \frac{\tilde{A} + \frac{a_u - 2a_\ell}{\sqrt{n}} \mathbb{1}_{n \times p}}{2(a_u - a_\ell)\sqrt{n}}. \quad (4)$$

Assumption 2.1 is originally imposed in Jiang et al. (2014) and ensures that the positivity and flux-preserving conditions are satisfied. By Lemma 2.1 in Jiang et al. (2014) if matrix \tilde{A} satisfies $\tilde{A}_{i,j} \in \frac{1}{\sqrt{n}}[a_\ell, a_u]$ for all entries, then the sensing matrix A will satisfy the physical constraints (2) and (3).

Assumption 2.2. For all $u \in \mathbb{R}^p$, $\|u\|_0 \leq 2\tilde{K}$, there exists a constant $\delta_{\tilde{K}} \equiv \delta_{\tilde{K}}(n, p) > 0$ such that:

$$\|\tilde{A}Du\|_2^2 \leq (1 + \delta_{\tilde{K}})\|u\|_2^2,$$

where \tilde{K} is a constant satisfies $\tilde{K} = O(R_q(\frac{\log p}{T})^{-\frac{q}{2}})$, without loss of generality we can assume \tilde{K} is a integer.

Assumption 2.2 is an upper restricted eigenvalue condition similar to that imposed in Jiang et al. (2014). The main difference is that we use a different sparse parameter \tilde{K} instead of s . In particular we set $\tilde{K} = O(R_q(\frac{\log p}{T})^{-\frac{q}{2}})$ which grows as T grows. This assumption will be used for both the lower bound and upper bound for the ℓ_1 -based method. As pointed out in Jiang et al. (2014), Assumption 2.2 holds if $n \geq C_0 \tilde{K} \log(p/\tilde{K})$ for a re-scaled Bernoulli ensemble matrix \tilde{A} with $P(\tilde{A}_{ij} = \frac{1}{\sqrt{n}}) = P(\tilde{A}_{ij} = -\frac{1}{\sqrt{n}}) = \frac{1}{2}$, with probability at least $1 - e^{-C_1 n}$ using results in Baraniuk et al. (2008).

Finally we define an s -sparse localization quantity also introduced in Jiang et al. (2014). The interaction between the orthonormal basis matrix D and the sparsity constraint has an effect on the lower bound which is captured by this s -sparse localization quantity:

Definition 2.1. λ_s is said to be the s -sparse localization quantity of a matrix X if

$$\lambda_s = \lambda_s(X) := \max_{v \in \{-1, 0, 1\}^p; \|v\|_0 = s} \|Xv\|_\infty.$$

Our minimax lower bound depends on $\lambda_k(\bar{D})$. Different scalings for the k -sparse localization constant $\lambda_k(\bar{D})$ with basis D for both Fourier and wavelet transforms are provided in Jiang et al. (2014). Now we present the minimax lower bound.

Theorem 1. *If $f^* \in \mathcal{F}_{p,q,D}$, $p \geq \max(260, R_q T^{\frac{q}{2}})$, Assumption 2.1 and 2.2 hold with $0 \leq \delta_{\tilde{K}} < 1$, let $\lambda_k = \lambda_k(\bar{D})$ be the k -sparse localization quantity of \bar{D} , then if $\tilde{K} = O(R_q(\frac{\log p}{T})^{-\frac{q}{2}})$, there exists a constant $C_L > 0$ that depends on a_u, a_ℓ and $\delta_{\tilde{K}}$ such that*

$$\min_f \max_{f^* \in \mathcal{F}_{p,q,D}} \|f - f^*\|_2^2 \geq C_L \max_{1 \leq k \leq \tilde{K}} \left\{ \min \left(\frac{k}{p^2 \lambda_k^2}, \frac{k \log \frac{p}{k}}{T} \right) \right\}$$

with probability greater than $\frac{1}{2}$.

Remarks:

- Note that in Theorem 1 when $k = \tilde{K} = O(R_q(\frac{\log p}{T})^{-\frac{q}{2}})$ and under condition $p \geq R_q T^{\frac{q}{2}}$, the second term $\frac{k \log \frac{p}{k}}{T}$ in the lower bound scales as $R_q(\frac{\log p}{T})^{1-\frac{q}{2}}$.

- In the case $q = 0$, the minimax rate is $\frac{s \log p}{T}$ as proven in Jiang et al. (2014). Note that if we set $s = s_q = R_q(\frac{\log p}{T})^{-q/2}$ leading to the overall rate $R_q(\frac{\log p}{T})^{1-q/2}$. This interpretation is consistent with the case of Gaussian linear models discussed in Raskutti et al. (2011).
- Note that like in the case $q = 0$ discussed in Jiang et al. (2014), the minimax lower bound depends on the intensity T and not the sample size n . This may initially seem counter-intuitive since the sample size n plays no role in the minimax lower bound. This phenomenon arises due to a combination of the signal-dependent and the flux-preserving constraint and we discuss in greater detail in Section 3.1.
- Values for λ_k are displayed in Table 1 of Jiang et al. (2014) for the discrete cosine transform (DCT), discrete Hadamard transform (DHT), and a discrete Haar wavelet basis (DWT). In particular for the DCT and DHT basis, $\lambda_k = \frac{\sqrt{2}k}{\sqrt{p}}$ and for DWT, $\lambda_k = \frac{1}{\sqrt{2}-1}$. Therefore the first term $\frac{k}{p^2 \lambda_k^2}$ in the lower bound will be $\frac{1}{pk}$ for DCT and DHT, and $\frac{k}{p^2}$ for DWT. Since when $k = \tilde{K}$ the second term $\frac{k \log \frac{p}{k}}{T}$ scales as $R_q(\frac{\log p}{T})^{1-\frac{q}{2}}$, then for DCT and DHT basis, $R_q(\frac{\log p}{T})^{1-\frac{q}{2}}$ will be smaller than the first term $\frac{1}{p\tilde{K}}$ when $p = o(T^{1-q})$, and for DWT, $R_q(\frac{\log p}{T})^{1-\frac{q}{2}}$ will be smaller than $\frac{\tilde{K}}{p^2}$ when $p = o(\sqrt{T})$.
- The proof in the case of ℓ_q -ball sparsity is more challenging than the strictly sparse case since the ℓ_q -ball is a compact set and we need to construct a packing set for the intersection of the ℓ_q -ball with the physical constraints on f^* . Our packing set is based on a combination of the hypercube construction provided in Kühn (2001) along with the construction in Jiang et al. (2014) which incorporates the positivity and flux-preserving constraints. Further details are provided in Section 4.1.

2.2 ℓ_1 -based Method

In this section we present an ℓ_2 -error upper bound for an ℓ_1 -based estimator by adapting existing results and techniques in Negahban et al. (2012) to our Poisson inverse problem setting. The estimator we consider is the standard Lasso estimator:

$$\hat{\theta}_{\lambda_n} \in \arg \min_{\theta \in \mathbb{R}^p, \theta_1 = \frac{1}{\sqrt{p}}} \frac{1}{n} \left\| \frac{n}{T} (y - TAD\theta) \right\|_2^2 + \lambda_n \|\theta\|_1$$

or equivalently

$$\hat{\theta}_{\lambda_n} \in \arg \min_{\theta \in \mathbb{R}^p, \theta_1 = \frac{1}{\sqrt{p}}} \frac{n}{T^2} \|y - TAD\theta\|_2^2 + \lambda_n \|\theta\|_1, \quad (5)$$

where $\lambda_n > 0$ is the regularization parameter. Next we introduce two further assumptions.

Assumption 2.3. There exists a constant $M_1 > 0$ such that

$$8n \log n < T < M_1 \left(\frac{n}{R_q \log p} \right)^{\frac{2}{q}} \log p \log n.$$

This assumption provides an upper and lower bound for the intensity T which we require when we prove the concentration bounds for the upper bounds. The lower bound for T is similar to the *high intensity* setting described in Jiang et al. (2014). From another perspective, Assumption 2.3 controls the signal-to-noise ratio which is controlled by the ratio $\frac{T}{n}$.

Assumption 2.4. There are strictly positive constants (k_1, k_2) that depend on a_u, a_ℓ such that

$$\sqrt{n} \|A\bar{D}x\|_2 \geq k_1 \|x\|_2 - k_2 \sqrt{\frac{\log p}{n}} \|x\|_1, \quad \forall x \in \mathbb{R}^{p-1}.$$

Assumption 2.4 is equivalent to the so-called restricted eigenvalue condition (see e.g. Bickel et al. (2008); Raskutti et al. (2010); van de Geer and Bühlmann (2009)) for the matrix $A\bar{D}$. This assumption holds for many appropriate choices of A , as we show in Theorem 3 in Section 2.3.

The upper bound is as follows:

Theorem 2. Suppose $f^* \in \mathcal{F}_{p,q,D}$ and Assumption 2.1, 2.2, 2.3 and 2.4 hold with $0 \leq \delta_{\tilde{K}} < 1$. Then there exists a constant $C_U > 0$ that depends on a_u, a_ℓ and $\delta_{\tilde{K}}$ such that

$$\|\hat{f}_{\lambda_n} - f^*\|_2^2 \leq C_U R_q \left(\frac{\log n \log p}{T} \right)^{1-\frac{q}{2}}$$

with probability at least $1 - \frac{2}{p-1}$.

Remarks:

- Note that the upper bound matches our minimax lower bound up to a $\log n$ factor. The $\log n$ factor is an artifact of our analysis since one of the important steps of

our proof requires us showing that linear combinations of Poisson random variables have sharp concentration bounds and the $\log n$ factor arises here in showing that the total counts are bounded with high probability. To prove that the total counts are bounded, we use Bobkov and Ledoux (1998) and develop our own two-sided version in combination with the union bound. The lower bound on T from Assumption 2.3 is needed for our concentration. Once we have bounded the Poisson counts, we exploit classical concentration bounds for linear combinations of bounded random variables using results in Hoeffding (1963).

- Aside from the $\log n$ factor, the upper and lower bounds match with a mean-squared error rate $R_q(\frac{\log p}{T})^{1-q/2}$. Once again note the dependence on T and not n , so the only place that n enters is through Assumption 2.3.

2.3 Restricted Eigenvalue Condition

In this section we show that Assumption 2.4 is satisfied with high probability for many choices of random matrices A under the appropriate scaling. In particular we show that the restricted eigenvalue condition is satisfied by matrices A with independent sub-Gaussian entries which include independent Bernoulli ensembles that also satisfy our flux-preserving and positivity constraints.

To characterize the sub-Gaussian parameter of a random variable, we define the Orlicz norm $\|\cdot\|_{\psi_2}$ for a random variable $X \in \mathbb{R}$ as follows:

$$\|X\|_{\psi_2} := \inf\{t : \mathbb{E} \exp(X^2/t^2) \leq 2\}.$$

The Orlicz norm as defined above is known to represent the sub-Gaussian parameter of a random variable. For example if $X \sim \mathcal{N}(0, \sigma^2)$, $\|X\|_{\psi_2} = \sigma$. Now we provide a definition of isotropic random vectors introduced in Mendelson et al. (2007) and Zhou (2009).

Definition 2.2 (Zhou (2009), Definition 1.3). Let Y be a random vector in \mathbb{R}^p ; Y is called isotropic if for every $y \in \mathbb{R}^p$, $\mathbb{E}|\langle Y, y \rangle|^2 = \|y\|_2^2$, and is ψ_2 with a constant α if for every $y \in \mathbb{R}^p$:

$$\|\langle Y, y \rangle\|_{\psi_2} \leq \alpha \|y\|_2.$$

Important examples of isotropic vectors are the Gaussian random vector $Y = (h_1, \dots, h_p)$ where $h_i, \forall i$ are independent $N(0, 1)$ random variables where $\alpha = 1$, and random vectors $Y = (\epsilon_1, \dots, \epsilon_p)$ where $\epsilon_i, \forall i$ are independent, symmetric ± 1 Bernoulli random variables also with $\alpha = 1$. Now we are ready to state the main theorem for this section:

Theorem 3. *There exists positive constants c', c'' for which the following holds. Let μ to be an isotropic ψ_2 probability measure with constant $\alpha \geq 1$. Let X_1, \dots, X_n be independent, distributed according to μ and define $\Gamma = \sum_{i=1}^n \langle X_i, \cdot \rangle e_i$. Then with probability at least $1 - c' \exp(-c''n)$, for all $x \in \mathbb{R}^p$ we will have*

$$\frac{\|x\|_2}{4} - C_\alpha \sqrt{\frac{\log p}{n}} \|x\|_1 \leq \frac{\|\Gamma x\|_2}{\sqrt{n}},$$

where C_α is a positive constant only depends on α .

Remarks:

- Theorem 3 shows that the restricted eigenvalue condition holds for matrices with random sub-Gaussian entries which include both Gaussian and bounded random variables. The proof techniques are based on a combination of techniques from Raskutti et al. (2010) for random Gaussian matrices with techniques from Mendelson et al. (2007) for sub-Gaussian random variables. The proof is provided in Section 4.3.
- Based on this theorem, there are many choices of A which satisfy Assumption 2.4. In our particular context, we also require A to satisfy Assumption 2.1 so that it satisfies our physical constraints. Hence we require the entries of A to be bounded, and we provide a concrete example below.
- Theorem 3 is more general than the restricted eigenvalue condition for strictly sparse vectors proven by Zhou (2009). Our result easily adapts to weak ℓ_q -ball sparse vectors and in addition our result applies to any x that may be random which we address using a peeling argument in our proof.

To construct a random matrix A that satisfies the restricted eigenvalue condition and Assumption 2.1, let \tilde{A} have the following entries:

$$\mathbb{P}(\tilde{A}_{ij}) = \begin{cases} \frac{1}{2} & \tilde{A}_{ij} = -\sqrt{\frac{1}{n}} \\ \frac{1}{2} & \tilde{A}_{ij} = \sqrt{\frac{1}{n}}, \end{cases}$$

then $\sqrt{n}\tilde{A}$ will satisfy the conditions for Γ in Theorem 3. Since we want our result to apply after we apply an orthonormal basis D we use the following Lemma:

Lemma 1. *Let μ to be an isotropic ψ_2 probability measure with constant $\alpha \geq 1$. And let $X \in \mathbb{R}^p$ be distributed according to μ , then $X\bar{D} \in \mathbb{R}^{p-1}$ is distributed according to another isotropic ψ_2 probability measure μ' with some constant $\alpha' \geq 1$.*

Thus by Lemma 1, $\sqrt{n}\tilde{A}\bar{D}$ satisfies the restricted eigenvalue condition from Theorem 3 and with high probability:

$$\|\tilde{A}\bar{D}x\|_2 = \frac{\|\sqrt{n}\tilde{A}\bar{D}x\|_2}{\sqrt{n}} \geq \frac{\|x\|_2}{4} - C\sqrt{\frac{\log p}{n}}\|x\|_1, \quad \forall x \in \mathbb{R}^{p-1},$$

where $C > 0$ is some absolute constant. Note that by the construction of A in Assumption 2.1 and definition of \bar{D} we have

$$A\bar{D}x = \frac{\tilde{A} + \frac{a_u - 2a_\ell}{\sqrt{n}}\mathbb{1}_{n \times p}}{2(a_u - a_\ell)\sqrt{n}}\bar{D}x = \frac{\tilde{A}\bar{D}x}{2(a_u - a_\ell)\sqrt{n}}.$$

Then

$$\sqrt{n}\|A\bar{D}x\|_2 \geq \frac{\|x\|_2}{8(a_u - a_\ell)} - \frac{C}{2(a_u - a_\ell)}\sqrt{\frac{\log p}{n}}\|x\|_1, \quad \forall x \in \mathbb{R}^{p-1},$$

which satisfies Assumption 2.4.

3 Discussion

In this section, we discuss some of the consequences and intuition for our three main results. In particular we discuss the dependence of the rates on n and T and discuss connections to the results for the Gaussian linear model in Raskutti et al. (2011), how the results for the ℓ_q case relate to the strictly sparse case developed in Jiang et al. (2014) and finally we compare our upper bound to the upper bounds developed in the recent work of Jiang et al. (2015) based on the weighted Lasso.

3.1 Dependence on T and n

One of the interesting and perhaps surprising aspects about both the upper and lower bounds is that they depend explicitly on the the intensity T and not on the sample size n , aside

from the conditions on design and the $\log n$ factor. This phenomenon also occurred in the strictly sparse case in Jiang et al. (2014) where the rate is $\frac{s \log p}{T}$. To understand this, we relate our rate of $R_q(\frac{\log p}{T})^{1-q/2}$ to the earlier results developed in Raskutti et al. (2011) for the Gaussian linear model and see how the signal-dependent noise and physical constraints ensure the minimax rate depends on T and not n .

In the Gaussian linear model under the ℓ_q -ball constraint studied in Raskutti et al. (2011), we have

$$y = \bar{A}f^* + w,$$

where $y \in \mathbb{R}^n$, $\bar{A} \in \mathbb{R}^{n \times p}$ with $p > n$ and $w \sim N(0, \sigma^2 I_{n \times n})$, and we have the constraint $\|f^*\|_q^q \leq R_q$ with $0 < q \leq 1$. Raskutti et al. (2011) shows that the minimax rate is:

$$\min_{\hat{f}} \max_{f^* \in \mathbb{B}_q(R_q)} \|\hat{f} - f^*\|_2^2 \asymp R_q \left(\frac{\sigma^2 \log p}{n} \right)^{1-\frac{q}{2}},$$

with high probability. In particular, take note of the role of σ^2 in the minimax rate. Later work by Negahban et al. (2012) proves that the Lasso estimator achieves this minimax rate. We will show how the dependence of the scaling on T and not n follows from our comparison to the rates for the Gaussian linear model and the impact of σ^2 .

For our Poisson inverse problem we can express the model as follows:

$$y \sim \text{Poisson}(TAf^*),$$

which can be expressed equivalently

$$y = TAf^* + \omega,$$

or

$$y_i = T \sum_{j=1}^p A_{ij} f_j^* + \omega_i, \quad 1 \leq i \leq n,$$

where

$$\mathbb{E}(y_i) = T \sum_{j=1}^p A_{ij} f_j^* = \frac{\alpha T}{n}, \quad 1 \leq i \leq n,$$

and

$$\mathbb{E}(\omega_i) = 0, \quad \text{Var}(\omega_i) = \text{Var}(y_i) = \mathbb{E}(y_i) = \frac{\alpha T}{n}, \quad 1 \leq i \leq n,$$

where $\frac{1}{2} \leq \alpha \leq 1$ by Lemma 4 in section 4.1. Note that this scaling of A follows from the flux-preserving constraint $\|Af\|_1 \leq \|f\|_1$. Since we have $\mathbb{E}(y_i)$ scaling as $\frac{T}{n}$, to ensure that the mean of our observations has the same scaling as in the Gaussian linear model independent of n and T , we consider the normalized responses:

$$\tilde{y}_i = \frac{n}{T} y_i = n \sum_{j=1}^p A_{ij} f_j^* + \frac{n}{T} \omega_i, \quad 1 \leq i \leq n,$$

where now $\mathbb{E}(\tilde{y}_i) = \alpha$ has a scaling independent of n and T , by defining $\tilde{\omega}_i = \frac{n}{T} \omega_i$ we also have

$$\mathbb{E}(\tilde{\omega}_i) = 0 \quad \text{and} \quad \text{Var}(\tilde{\omega}_i) = \frac{n^2}{T^2} \frac{\alpha T}{n} = \frac{\alpha n}{T}.$$

Hence the combination of the signal-dependent noise and the flux-preserving constraint mean that we have σ^2 scaling as $\frac{n}{T}$ in the appropriately normalized model.

Recall that for the Gaussian model, the minimax rate is $R_q(\frac{\sigma^2 \log p}{n})^{1-q/2}$, and if we replace σ^2 by $\frac{\alpha n}{T}$ we will get the minimax rate scaling as $R_q(\frac{\alpha \log p}{T})^{1-q/2}$, which is consistent with our minimax lower bound. Hence, if we want to relate our physically constrained Poisson model to the Gaussian linear model, we need to consider a model with variance $\sigma^2 \propto \frac{n}{T}$.

This observation that the minimax rate depends on the signal intensity T rather than the sample size n was also made in the recent work of Jiang et al. (2014). Our analysis shows that this observation carries over to the ℓ_q -ball setting. The caveat is that n is required to be sufficiently large to ensure that the restricted eigenvalue is satisfied which is also required in the strictly sparse case.

3.2 Comparison to related results

For the strictly sparse case studied by Jiang et al. (2014), the minimax rate scales as $\frac{s \log p}{T}$ whereas in the weakly sparse case in this paper, the minimax rate scales as $R_q(\frac{\log p}{T})^{1-q/2}$. Another way to interpret our result for the ℓ_q -ball case is that the minimax rate scales as $\frac{s_q \log p}{T}$ where $s_q = O(R_q(\frac{\log p}{T})^{-q/2})$. This can be explained in terms of a bias-variance tradeoff to determine how many co-ordinates of f^* should be included in the model and using the ℓ_q -ball constraint, selecting $s_q = O(R_q(\frac{\log p}{T})^{-q/2})$ with largest magnitude optimizes the bias-variance tradeoff to minimize the mean-squared error. This interpretation is used at several

points in the proofs of both the minimax lower bound and upper bound. This phenomenon was also observed in the Gaussian linear model case in Raskutti et al. (2011).

Another recent related work is by Jiang et al. (2015) which provides analysis for a weighted Lasso estimator. In Jiang et al. (2015) sparse Poisson inverse problems under the model $Y \sim \text{Poisson}(Af^*)$ are discussed and Jiang et al. (2015) provides a weighted Lasso estimator \hat{f}^{WL} based on the minimizer of the following optimization problem:

$$\hat{f}^{WL} \in \arg \min_f \|\tilde{Y} - \tilde{A}f\|_2^2 + \gamma \sum_{j=1}^p d_j |f_j|,$$

where \tilde{Y} and \tilde{A} are shifted and scaled versions of Y and A , $\gamma > 2$ is a constant and positive weights $(d_j)_{j=1}^p$ are chosen in a specific way to minimize mean-squared error. In the case where all the weights are the same $d_j = \lambda_n$ for all j which corresponds to the ordinary Lasso estimator we analyze in this paper. We summarize their result and show that it is sub-optimal for ℓ_q -balls. To be clear, the focus of the results in Jiang et al. (2015) is the strictly sparse case in a number of more general settings than this paper where they achieve optimal or near-optimal mean-squared error. However, they do have a result for approximately sparse models which is not focussed specifically on the ℓ_q -ball sparsity setting. To summarize their result, they introduce a bias term:

$$B_s := \max\{\|\tilde{A}(f^* - f_s^*)\|_2^2, \|f^* - f_s^*\|_1\},$$

where $s > 0$ is an integer and f_s^* is the best s -sparse approximation to f^* , then they state that

$$\|f^* - \hat{f}^{WL}\|_2^2 \preceq B_s + \lambda^2 s + (1 + 1/\lambda)^2 B_s^2. \quad (6)$$

It must be pointed out that Jiang et al. (2015) analyze a broader choice of weights $(d_j)_{j=1}^p$ but we were unable to find different choices of weights that provided a sharper upper bound in our ℓ_q -ball context. As discussed in Jiang et al. (2015), their analysis yields the optimal rates (up to $\log n$ factors) if f^* is s -sparse and the bias term $B_s = 0$. In the case of ℓ_q -ball sparsity as discussed earlier an appropriate choice for s is $s = s_q = O(R_q(\frac{\log p}{T})^{-\frac{q}{2}})$ and $\lambda_n = O(\sqrt{\frac{\log p}{T}})$, for the B_s term, $\|f^* - f_s^*\|_1$ will be bounded by $R_q \lambda_n^{1-q}$ by inequality (34). By replacing these terms in (6), $\|f^* - \hat{f}^{WL}\|_2^2$ will be bounded by $O(R_q \lambda_n^{1-q} + R_q^2 \lambda_n^{-2q})$, which scales as $R_q(\frac{\log p}{T})^{\frac{1-q}{2}} + R_q^2(\frac{\log p}{T})^{-q}$, this bound is clearly sub-optimal when $T > \log p$ since the upper bound provided in our result scales as $R_q(\frac{\log p}{T})^{1-\frac{q}{2}}$.

4 Proofs

In this section we provide the proofs for our three main results. We defer the more technical steps to the appendix.

4.1 Proof of Theorem 1

The proof for the lower bound uses a combination of standard information-theoretic techniques involving Fano's inequality, and the explicit construction of a packing set that satisfies the ℓ_q -ball constraint and our other physical constraints. In particular, the proof involves constructing a packing set for $\mathcal{F}_{p,q,D}$ and then applying the generalized Fano method to the packing set (see Han and Verdu (1994), Ibragimov and Has'minskii (1981) and Yang and Barron (1999) for details). Constructing the packing set is the main challenge and novelty in the proof. Our packing set is based on a constrained hypercube construction in Jiang et al. (2014) along with the hyper-cube construction for ℓ_q -balls in Kühn (2001).

Proof. We begin our proof by constructing a packing set for $\mathcal{F}_{p,q,D}$. For $1 \leq k \leq \tilde{K}$ let $\mathcal{H}_k = \{\beta \in \{-1, 0, +1\}^{p-1} : \|\beta\|_0 = k\}$. From [Raskutti et al. (2011), Lemma 4] we can find a subset $\tilde{\mathcal{H}}_k \subseteq \mathcal{H}_k$ with cardinality $|\tilde{\mathcal{H}}_k| \geq \exp(\frac{k}{2} \log \frac{p-\frac{k}{2}-1}{k})$ such that the Hamming distance $\rho_H(\beta, \beta') \geq \frac{k}{2}$ for all $\beta, \beta' \in \tilde{\mathcal{H}}_k$. Then note that $(\frac{R_q}{k})^{\frac{1}{q}} \mathcal{H}_k = \{\beta \in \{-(\frac{R_q}{k})^{\frac{1}{q}}, 0, +(\frac{R_q}{k})^{\frac{1}{q}}\}^{p-1} : \|\beta\|_q^q = R_q\}$. Now consider the re-scaled hypercube $\tilde{\mathcal{H}}_k$ by α_k with $0 < \alpha_k \leq (\frac{R_q}{k})^{\frac{1}{q}}$, we define: $\mathcal{H}_{k,\alpha_k} = \{\theta \in \mathbb{R}^p : \theta = [1/\sqrt{p}, \alpha_k \beta^\top]^\top, \beta \in \tilde{\mathcal{H}}_k\}$, let $\eta_{\alpha_k}^2 = \frac{k}{2} \alpha_k^2$, then \mathcal{H}_{k,α_k} is a η_{α_k} -packing set for $\mathcal{F}_{p,q,D}$ in the ℓ_2 norm. To contrast with the packing set used in Jiang et al. (2014), we require the extra constraint that each vertex has value at most $(\frac{R_q}{k})^{\frac{1}{q}}$ to ensure $\|\beta\|_q^q \leq R_q$.

The following lemma shows useful properties of this packing set.

Lemma 2 (Jiang et al. (2014), Lemma 4.1). *For $1 \leq k \leq \tilde{K}$, let $\lambda_k = \lambda_k(\bar{D})$. Then the packing sets \mathcal{H}_{k,α_k} with $0 < \alpha_k \leq \frac{1}{p\lambda_k}$ have the following properties:*

1. *The ℓ_2 distance between any two points θ and θ' in \mathcal{H}_{k,α_k} is bounded:*

$$\eta_{\alpha_k}^2 \leq \|\theta - \theta'\|_2^2 \leq 8\eta_{\alpha_k}^2.$$

2. For any $\theta \in \mathcal{H}_{k,\alpha_k}$, the corresponding $f = D\theta$ satisfies:

$$f_i \geq 0, \forall i \in \{1, 2, \dots, p\}, \quad \text{and} \quad \|f\|_1 = 1.$$

3. The size of the packing set

$$|\mathcal{H}_{k,\alpha_k}| \geq \exp\left(\frac{k}{2} \log \frac{p - \frac{k}{2} - 1}{k}\right).$$

The proof for this lemma can be found in Jiang et al. (2014).

For convenience we define the matrix $\Phi \triangleq AD$ and then $\Phi\theta = Af$ since $f = D\theta$. Next we will apply the generalized Fano method to the packing set, these techniques are developed in Han and Verdu (1994), Ibragimov and Has'minskii (1981) and Yang and Barron (1999). Define M_k to be the cardinality of the set \mathcal{H}_{k,α_k} , and the elements in \mathcal{H}_{k,α_k} can be denoted as $\{\theta^1, \dots, \theta^{M_k}\}$. Let $\tilde{\Theta} \in \mathbb{R}^p$ be a random vector drawn from a uniform distribution over the packing set $\{\theta^1, \dots, \theta^{M_k}\}$. Further let $\tilde{\theta} = \arg \min_{\theta \in \mathcal{H}_{k,\alpha_k}} \|\theta - D^\top f\|_2$. Then since D is an orthonormal basis we can bound the minimax estimation error according to Yang and Barron (1999):

$$P(\min_f \max_{f^* \in \mathcal{F}_{p,q,D}} \|f - f^*\|_2^2 \geq \frac{\eta_{\alpha_k}^2}{4}) \geq \min_{\tilde{\theta}} \mathbb{P}[\tilde{\theta} \neq \tilde{\Theta}]. \quad (7)$$

Applying Fano's inequality yields the following lower bound:

$$\mathbb{P}[\tilde{\theta} \neq \tilde{\Theta}] \geq 1 - \frac{I(y; \tilde{\Theta}) + \log 2}{\log M_k}, \quad (8)$$

where $y|\tilde{\Theta} \sim \text{Poisson}(T\Phi\tilde{\Theta})$ and $I(y; \tilde{\Theta})$ is the mutual information between random variable y and $\tilde{\Theta}$. Then from Han and Verdu (1994) we have

$$I(y; \tilde{\Theta}) \leq \frac{1}{\binom{M_k}{2}} \sum_{i,j=1,\dots,M_k, i \neq j} \text{KL}(p(y|T\Phi\theta^i) \| p(y|T\Phi\theta^j)), \quad (9)$$

where $\text{KL}(p_1 \| p_2)$ is the Kullback-Leibler (KL) divergence between distributions p_1 and p_2 . We will use the following lemma to bound KL divergence of Poisson distributions in terms of the squared ℓ_2 -distance.

Lemma 3 (Jiang et al. (2014), Lemma 4.2). *Let $p(y|u)$ denote the vector Poisson distribution with mean parameter $\mu \in \mathbb{R}_+^n$. For $\mu_1, \mu_2 \in \mathbb{R}_+^n$, if there exists some value $c > 0$ such that $\mu_2 \succeq c\mathbb{1}_{n \times 1}$, then the following holds:*

$$\text{KL}(p(y|\mu_1) \| p(y|\mu_2)) \leq \frac{1}{c} \|\mu_1 - \mu_2\|_2^2.$$

The following lemma shows that entries in Af^* are bounded between $\frac{1}{2n}$ and $\frac{1}{n}$ under Assumption 2.1:

Lemma 4 (Jiang et al. (2014), Lemma 4.3). *If the sensing matrix A satisfies Assumption 2.1, then for all nonnegative f with $\|f\|_1 = 1$, we have:*

$$\frac{1}{2n}\mathbb{1}_{n \times 1} \preceq Af \preceq \frac{1}{n}\mathbb{1}_{n \times 1}.$$

The proofs for Lemmas 3 and 4 can be found in Jiang et al. (2014). By Lemma 4 we have $\Phi\theta^j = AD\theta^j \succeq \frac{1}{2n}\mathbb{1}_{n \times 1}$ and then it follows that $T\Phi\theta^j \succeq \frac{T}{2n}\mathbb{1}_{n \times 1}$. Then from Lemma 3 we can bound the KL divergence between $p(y|T\Phi\theta^i)$ and $p(y|T\Phi\theta^j)$ as follows:

$$\text{KL}(p(y|T\Phi\theta^i)||p(y|T\Phi\theta^j)) \leq \frac{2n}{T}\|T\Phi(\theta^i - \theta^j)\|_2^2 = 2nT\|\Phi(\theta^i - \theta^j)\|_2^2. \quad (10)$$

By Assumption 2.1 and 2.2 if we denote $f^i = D\theta^i$, $f^j = D\theta^j$, then

$$\begin{aligned} \|\Phi(\theta^i - \theta^j)\|_2^2 &= \|A(f^i - f^j)\|_2^2 \\ &= \left\| \frac{1}{2(a_u - a_\ell)\sqrt{n}} \tilde{A}D(\theta^i - \theta^j) \right\|_2^2 \\ &\leq \frac{1 + \delta_{\tilde{K}}}{4(a_u - a_\ell)^2 n} \|\theta^i - \theta^j\|_2^2. \end{aligned} \quad (11)$$

Since $\|\theta^i - \theta^j\|_2^2 \leq 8\eta_{\alpha_k}^2$ by Lemma 2, we further have

$$\|\Phi(\theta^i - \theta^j)\|_2^2 \leq \frac{2(1 + \delta_{\tilde{K}})}{(a_u - a_\ell)^2 n} \eta_{\alpha_k}^2. \quad (12)$$

Then by combining (10) and (12) we have:

$$\text{KL}(p(y|T\Phi\theta^i)||p(y|T\Phi\theta^j)) \leq 2nT\|\Phi(\theta^i - \theta^j)\|_2^2 \leq \frac{4(1 + \delta_{\tilde{K}})T}{(a_u - a_\ell)^2} \eta_{\alpha_k}^2. \quad (13)$$

Then the mutual information can be bounded by using (9) and (13)

$$\begin{aligned} I(y; \tilde{\Theta}) &\leq \frac{1}{\binom{M_k}{2}} \sum_{i \neq j} \text{KL}(p(y|T\Phi\theta^i)||p(y|T\Phi\theta^j)) \\ &\leq \max_{i \neq j} \text{KL}(p(y|T\Phi\theta^i)||p(y|T\Phi\theta^j)) \\ &\leq \frac{4(1 + \delta_{\tilde{K}})T}{(a_u - a_\ell)^2} \eta_{\alpha_k}^2. \end{aligned} \quad (14)$$

Using (8) and the lower bound for M_k we have

$$\mathbb{P}[\tilde{\theta} \neq \tilde{\Theta}] \geq 1 - \frac{I(y; \tilde{\Theta}) + \log 2}{\log M_k} \geq 1 - \frac{\frac{4(1+\delta_{\tilde{K}})T}{(a_u - a_\ell)^2} \eta_{\alpha_k}^2 + \log 2}{\frac{k}{2} \log \frac{p - \frac{k}{2} - 1}{k}}. \quad (15)$$

Next we will show that the probability in (15) is bounded by the constant $1/2$. This constant is guaranteed if the following two inequalities are true:

$$\frac{k}{2} \log \frac{p - \frac{k}{2} - 1}{k} \geq 4 \log 2, \quad (16)$$

$$\frac{k}{2} \log \frac{p - \frac{k}{2} - 1}{k} \geq \frac{16(1 + \delta_{\tilde{K}})T}{(a_u - a_\ell)^2} \eta_{\alpha_k}^2. \quad (17)$$

For the first inequality (16) if $k = 1$,

$$\frac{k}{2} \log \frac{p - \frac{k}{2} - 1}{k} = \frac{1}{2} \log \left(p - \frac{3}{2} \right) \geq \frac{1}{2} \log \left(\frac{515}{2} - \frac{3}{2} \right) = 4 \log 2,$$

where the inequality is a result of $p \geq 260$. And if $k \geq 2$,

$$\begin{aligned} \frac{k}{2} \log \frac{p - \frac{k}{2} - 1}{k} &\geq \log \frac{p - \frac{k}{2} - 1}{k} \\ &\geq \log \frac{p - \frac{\tilde{K}}{2} - 1}{\tilde{K}} \\ &\geq \log \frac{(\frac{33\tilde{K}}{2} + 1) - (\frac{\tilde{K}}{2} + 1)}{\tilde{K}} \\ &= 4 \log 2, \end{aligned}$$

where the inequality is the result of $p \geq R_q T^{\frac{4}{2}} \geq \frac{33\tilde{K}}{2} + 1$ when p is big enough. For the second inequality (17) we need:

$$\frac{k}{32} \log \frac{p - \frac{k}{2} - 1}{k} \geq \frac{(1 + \delta_{\tilde{K}})T}{(a_u - a_\ell)^2} \eta_{\alpha_k}^2,$$

which leads to

$$\eta_{\alpha_k}^2 \leq \frac{(a_u - a_\ell)^2 k}{32(1 + \delta_{\tilde{K}})T} \log \frac{p - \frac{k}{2} - 1}{k}.$$

Since for Lemma 2 we require that $0 < \alpha_k \leq \frac{1}{p\lambda_k}$, also recall the extra constraint that $\alpha_k \leq (\frac{R_q}{k})^{\frac{1}{q}}$, thus we have:

$$\eta_{\alpha_k}^2 = \min \left(\frac{k}{2} \left(\frac{1}{p\lambda_k} \right)^2, \frac{(a_u - a_\ell)^2 k}{32(1 + \delta_{\tilde{K}})T} \log \frac{p - \frac{k}{2} - 1}{k}, \frac{k}{2} \left(\frac{R_q}{k} \right)^{\frac{2}{q}} \right).$$

Then with probability greater than $\frac{1}{2}$ we have

$$\begin{aligned} \min_f \max_{f^* \in \mathcal{F}_{p,q,D}} \|f - f^*\|_2^2 &\geq \frac{\eta_{\alpha_k}^2}{4} \\ &= \min \left(\frac{k}{8} \left(\frac{1}{p\lambda_k} \right)^2, \frac{(a_u - a_\ell)^2 k}{128(1 + \delta_{\tilde{K}})T} \log \frac{p - \frac{k}{2} - 1}{k}, \frac{k}{8} \left(\frac{R_q}{k} \right)^{\frac{2}{q}} \right). \end{aligned} \quad (18)$$

In order to further simplify (18), note that $1 \leq k \leq \tilde{K} = O(R_q(\frac{\log p}{T})^{-\frac{q}{2}})$, we then have

$$k \leq O(R_q(\frac{\log p}{T})^{-\frac{q}{2}}), \quad (19)$$

from (19) there exists some absolute constant $C_1 > 0$ such that

$$\frac{(a_u - a_\ell)^2 k}{128(1 + \delta_{\tilde{K}})T} \log \frac{p - \frac{k}{2} - 1}{k} \leq C_1 \frac{k}{8} \left(\frac{R_q}{k} \right)^{\frac{2}{q}}. \quad (20)$$

Thus by (18) and (20)

$$\min_f \max_{f^* \in \mathcal{F}_{p,q,D}} \|f - f^*\|_2^2 \geq \min \left(\frac{k}{8} \left(\frac{1}{p\lambda_k} \right)^2, \frac{(a_u - a_\ell)^2 k}{128 \max\{1, C_1\}(1 + \delta_{\tilde{K}})T} \log \frac{p - \frac{k}{2} - 1}{k} \right). \quad (21)$$

Since (21) holds for all $1 \leq k \leq \tilde{K}$, there exists a constant $C_L > 0$ such that

$$\min_f \max_{f^* \in \mathcal{F}_{p,q,D}} \|f - f^*\|_2^2 \geq C_L \max_{1 \leq k \leq \tilde{K}} \left\{ \min \left(\frac{k}{p^2 \lambda_k^2}, \frac{k \log \frac{p}{k}}{T} \right) \right\}$$

with probability greater than $\frac{1}{2}$. □

4.2 Proof for Theorem 2

The proof for the upper bound involves direct analysis of the lasso estimator defined in (5). Our analysis follows standard steps for analysis of regularized M-estimators (see Bickel et al. (2009); Negahban et al. (2012); van de Geer (2000)) along with addressing two challenges specific to this setting: (1) we use concentration bounds for linear combination of Poisson random variables and how they are used to determine a λ_n ; (2) use Assumption 2.3 to show that matrix $A\bar{D}$ satisfies the restricted eigenvalue condition and satisfies the physical constraints.

Proof. From (5) in Section 2.2 we know $\hat{\theta}_{\lambda_n}$ is a solution to the following problem:

$$\hat{\theta}_{\lambda_n} \in \arg \min_{\theta \in \mathbb{R}^p, \theta_1 = \frac{1}{\sqrt{p}}} \frac{n}{T^2} \|y - TAD\theta\|_2^2 + \lambda_n \|\theta\|_1, \quad (22)$$

with $(\hat{\theta}_{\lambda_n})_1 = \frac{1}{\sqrt{p}}$. Since θ^* satisfies the constraint that $\theta^* \in \mathbb{R}^p$ and $\theta_1^* = \frac{1}{\sqrt{p}}$, we have the following basic inequality

$$\frac{n}{T^2} \|y - TAD\hat{\theta}_{\lambda_n}\|_2^2 + \lambda_n \|\hat{\theta}_{\lambda_n}\|_1 \leq \frac{n}{T^2} \|y - TAD\theta^*\|_2^2 + \lambda_n \|\theta^*\|_1.$$

Hence

$$\frac{n}{T^2} \|TAD(\theta^* - \hat{\theta}_{\lambda_n})\|_2^2 \leq \frac{2n}{T^2} |(y - TAD\theta^*)^\top TAD(\theta^* - \hat{\theta}_{\lambda_n})| + \lambda_n (\|\theta^*\|_1 - \|\hat{\theta}_{\lambda_n}\|_1),$$

and then

$$n \|AD(\theta^* - \hat{\theta}_{\lambda_n})\|_2^2 \leq \frac{2n}{T} |(y - TAD\theta^*)^\top AD(\theta^* - \hat{\theta}_{\lambda_n})| + \lambda_n (\|\theta^*\|_1 - \|\hat{\theta}_{\lambda_n}\|_1). \quad (23)$$

Note that $(\theta^*)_1 = (\hat{\theta}_{\lambda_n})_1 = \frac{1}{\sqrt{p}}$, thus it is reasonable to define the error vector $\hat{\Delta} = \bar{\theta}^* - \bar{\hat{\theta}}_{\lambda_n} \in \mathbb{R}^{p-1}$, where $\bar{\theta}^* = [\theta_2^*, \dots, \theta_p^*]^\top \in \mathbb{R}^{p-1}$ and $\bar{\hat{\theta}}_{\lambda_n} = [(\hat{\theta}_{\lambda_n})_2, \dots, (\hat{\theta}_{\lambda_n})_p]^\top \in \mathbb{R}^{p-1}$. Then (23) can be reduced to:

$$\begin{aligned} n \|A\bar{D}\hat{\Delta}\|_2^2 &\leq \frac{2n}{T} |(y - TAD\theta^*)^\top A\bar{D}\hat{\Delta}| + \lambda_n (\|\theta^*\|_1 - \|\hat{\theta}_{\lambda_n}\|_1) \\ &\leq \left\| \frac{2n}{T} (y - TAD\theta^*)^\top A\bar{D} \right\|_\infty \|\hat{\Delta}\|_1 + \lambda_n (\|\theta^*\|_1 - \|\hat{\theta}_{\lambda_n}\|_1), \end{aligned} \quad (24)$$

where $\bar{D} = [d_2, \dots, d_p] \in \mathbb{R}^{p \times (p-1)}$.

In order to associate the term $\|\theta^*\|_1 - \|\hat{\theta}_{\lambda_n}\|_1$ with $\hat{\Delta}$, we define a threshold parameter $\eta > 0$ and the threshold subset as follows:

$$S_\eta := \{j \in \{2, 3, \dots, p\} \mid |\theta_j^*| > \eta\}$$

and its complement

$$S_\eta^c := \{j \in \{2, 3, \dots, p\} \mid |\theta_j^*| \leq \eta\}.$$

Suppose u is a vector in \mathbb{R}^{p-1} , we will define $u_{S_\eta} \in \mathbb{R}^{p-1}$ as following:

$$(u_{S_\eta})_j = \begin{cases} u_j & \text{if } j+1 \in S_\eta \\ 0 & \text{if } j+1 \notin S_\eta \end{cases} \quad \text{for } 1 \leq j \leq p-1,$$

and $u_{S_\eta^c}$ is defined in a similar way. Now we show how to connect $\|\theta^*\|_1 - \|\hat{\theta}_{\lambda_n}\|_1$ with $\hat{\Delta}$. Note that

$$\|\theta^*\|_1 - \|\hat{\theta}_{\lambda_n}\|_1 = \|\bar{\theta}^*\|_1 - \|\bar{\hat{\theta}}_{\lambda_n}\|_1, \quad (25)$$

since $\theta_1^* = (\hat{\theta}_{\lambda_n})_1 = \frac{1}{\sqrt{p}}$. Then by using the triangle inequality we have

$$\begin{aligned} \|\bar{\hat{\theta}}_{\lambda_n}\|_1 = \|\bar{\theta}^* - \hat{\Delta}\|_1 &= \|\bar{\theta}_{S_\eta}^* + \bar{\theta}_{S_\eta^c}^* - \hat{\Delta}_{S_\eta} - \hat{\Delta}_{S_\eta^c}\|_1 \\ &\geq \|\bar{\theta}_{S_\eta}^* - \hat{\Delta}_{S_\eta^c}\|_1 - \|\bar{\theta}_{S_\eta^c}^*\|_1 - \|\hat{\Delta}_{S_\eta}\|_1 \\ &= \|\bar{\theta}_{S_\eta}^*\|_1 + \|\hat{\Delta}_{S_\eta}\|_1 - \|\bar{\theta}_{S_\eta^c}^*\|_1 - \|\hat{\Delta}_{S_\eta}\|_1. \end{aligned} \quad (26)$$

On the other hand we have $\|\bar{\theta}^*\|_1 \leq \|\bar{\theta}_{S_\eta}^*\|_1 + \|\bar{\theta}_{S_\eta^c}^*\|_1$. Thus by combining these two inequalities we have

$$\|\bar{\theta}^*\|_1 - \|\bar{\hat{\theta}}_{\lambda_n}\|_1 \leq \|\hat{\Delta}_{S_\eta}\|_1 - \|\hat{\Delta}_{S_\eta^c}\|_1 + 2\|\bar{\theta}_{S_\eta^c}^*\|_1.$$

Therefore by (25):

$$\|\theta^*\|_1 - \|\hat{\theta}_{\lambda_n}\|_1 \leq \|\hat{\Delta}_{S_\eta}\|_1 - \|\hat{\Delta}_{S_\eta^c}\|_1 + 2\|\bar{\theta}_{S_\eta^c}^*\|_1. \quad (27)$$

By using (27) in (24) we have

$$n\|A\bar{D}\hat{\Delta}\|_2^2 \leq \left\|\frac{2n}{T}(y - TAD\theta^*)^\top A\bar{D}\right\|_\infty \|\hat{\Delta}\|_1 + \lambda_n(\|\hat{\Delta}_{S_\eta}\|_1 - \|\hat{\Delta}_{S_\eta^c}\|_1 + 2\|\bar{\theta}_{S_\eta^c}^*\|_1). \quad (28)$$

Now we upper bound the $\|\cdot\|_\infty$ norm through the following Lemma:

Lemma 5. *Under Assumption 2.1, 2.2 and 2.3, with probability at least $1 - \frac{2}{p-1}$ that*

$$\left\|\frac{2n}{T}(y - TAD\theta^*)^\top A\bar{D}\right\|_\infty \leq \sqrt{\frac{512M \log n \log p}{T}},$$

where $M = \frac{1+\delta_{\bar{K}}}{4(a_u - a_\ell)^2}$.

The proof for Lemma 5 is deferred to the appendix.

Thus by setting $\lambda_n = 2\sqrt{\frac{512M \log n \log p}{T}}$ in (28) we have

$$\begin{aligned} n\|A\bar{D}\hat{\Delta}\|_2^2 &\leq \frac{\lambda_n}{2} \|\hat{\Delta}\|_1 + \lambda_n(\|\hat{\Delta}_{S_\eta}\|_1 - \|\hat{\Delta}_{S_\eta^c}\|_1 + 2\|\bar{\theta}_{S_\eta^c}^*\|_1) \\ &\leq \frac{\lambda_n}{2} (\|\hat{\Delta}_{S_\eta}\|_1 + \|\hat{\Delta}_{S_\eta^c}\|_1 + 2\|\hat{\Delta}_{S_\eta}\|_1 - 2\|\hat{\Delta}_{S_\eta^c}\|_1 + 4\|\bar{\theta}_{S_\eta^c}^*\|_1) \\ &= \frac{\lambda_n}{2} (3\|\hat{\Delta}_{S_\eta}\|_1 - \|\hat{\Delta}_{S_\eta^c}\|_1 + 4\|\bar{\theta}_{S_\eta^c}^*\|_1), \end{aligned} \quad (29)$$

where the second inequality follows from the triangle inequality. From (29) we can see that $0 \leq 3\|\hat{\Delta}_{S_\eta}\|_1 - \|\hat{\Delta}_{S_\eta^c}\|_1 + 4\|\bar{\theta}_{S_\eta^c}^*\|_1$, then the error vector $\hat{\Delta}$ should satisfy $\|\hat{\Delta}_{S_\eta^c}\|_1 \leq 3\|\hat{\Delta}_{S_\eta}\|_1 + 4\|\bar{\theta}_{S_\eta^c}^*\|_1$. The following lemma shows that $n\|A\bar{D}\hat{\Delta}\|_2^2$ is lower bounded for all $\hat{\Delta} \in \{\Delta \in \mathbb{R}^{p-1} \mid \|\Delta_{S_\eta^c}\|_1 \leq 3\|\Delta_{S_\eta}\|_1 + 4\|\bar{\theta}_{S_\eta^c}^*\|_1\}$:

Lemma 6. *Suppose Assumption 2.3 and 2.4 hold, for all $\hat{\Delta} \in \{\Delta \in \mathbb{R}^{p-1} \mid \|\Delta_{S_\eta^c}\|_1 \leq 3\|\Delta_{S_\eta}\|_1 + 4\|\bar{\theta}_{S_\eta^c}^*\|_1\}$ we have*

$$n\|A\bar{D}\hat{\Delta}\|_2^2 \geq c_1 k_1^2 \|\hat{\Delta}\|_2^2 - c_2 k_2^2 \frac{\log p}{n} \|\bar{\theta}_{S_\eta^c}^*\|_1^2,$$

where $c_1, c_2 > 0$ are some constants.

Once again the proof of Lemma 6 is deferred to the appendix.

By using Lemma 6 in (29) we have

$$\begin{aligned} c_1 k_1^2 \|\hat{\Delta}\|_2^2 - c_2 k_2^2 \frac{\log p}{n} \|\bar{\theta}_{S_\eta^c}^*\|_1^2 &\leq \frac{\lambda_n}{2} (3\|\hat{\Delta}_{S_\eta}\|_1 - \|\hat{\Delta}_{S_\eta^c}\|_1 + 4\|\bar{\theta}_{S_\eta^c}^*\|_1) \\ &\leq \frac{\lambda_n}{2} (3\|\hat{\Delta}_{S_\eta}\|_1 + 4\|\bar{\theta}_{S_\eta^c}^*\|_1). \end{aligned} \quad (30)$$

Then note that $\|\hat{\Delta}_{S_\eta}\|_1 \leq \sqrt{|S_\eta|} \|\hat{\Delta}_{S_\eta}\|_2 \leq \sqrt{|S_\eta|} \|\hat{\Delta}\|_2$, where $|S_\eta|$ is the cardinality of set S_η , then from (30) we have

$$c_1 k_1^2 \|\hat{\Delta}\|_2^2 - c_2 k_2^2 \frac{\log p}{n} \|\bar{\theta}_{S_\eta^c}^*\|_1^2 \leq \frac{\lambda_n}{2} (3\sqrt{|S_\eta|} \|\hat{\Delta}\|_2 + 4\|\bar{\theta}_{S_\eta^c}^*\|_1),$$

which implies that

$$c_1 k_1^2 \|\hat{\Delta}\|_2^2 - \frac{3\lambda_n \sqrt{|S_\eta|}}{2} \|\hat{\Delta}\|_2 - c_2 k_2^2 \frac{\log p}{n} \|\bar{\theta}_{S_\eta^c}^*\|_1^2 - 2\lambda_n \|\bar{\theta}_{S_\eta^c}^*\|_1 \leq 0. \quad (31)$$

Note that the left hand side of (31) can be seen as a quadratic form of $\|\hat{\Delta}\|_2$. Thus by solving this quadratic inequality for $\|\hat{\Delta}\|_2$ we have

$$\|\hat{\Delta}\|_2^2 \leq 9 \frac{\lambda_n^2}{c_1^2 k_1^4} |S_\eta| + \frac{1}{c_1 k_1^2} (2c_2 k_2^2 \frac{\log p}{n} \|\bar{\theta}_{S_\eta^c}^*\|_1^2 + 4\lambda_n \|\bar{\theta}_{S_\eta^c}^*\|_1).$$

Hence

$$\begin{aligned} \|\hat{f}_{\lambda_n} - f^*\|_2^2 &= \|D(\hat{\theta}_{\lambda_n} - \theta^*)\|_2^2 = \|\hat{\theta}_{\lambda_n} - \theta^*\|_2^2 \\ &= \|\bar{\theta}_{\lambda_n} - \bar{\theta}^*\|_2^2 = \|\hat{\Delta}\|_2^2 \\ &\leq 9 \frac{\lambda_n^2}{c_1^2 k_1^4} |S_\eta| + \frac{1}{c_1 k_1^2} (2c_2 k_2^2 \frac{\log p}{n} \|\bar{\theta}_{S_\eta^c}^*\|_1^2 + 4\lambda_n \|\bar{\theta}_{S_\eta^c}^*\|_1). \end{aligned} \quad (32)$$

Since

$$R_q \geq \sum_{j \in S_\eta} |\theta_j^*|^q \geq \eta^q |S_\eta|, \quad (33)$$

we have $|S_\eta| \leq \eta^{-q} R_q$. On the other hand

$$\|\bar{\theta}_{S_\eta^c}^*\|_1 = \sum_{j \in S_\eta^c} |\bar{\theta}_j^*| = \sum_{j \in S_\eta^c} |\bar{\theta}_j^*|^q |\bar{\theta}_j^*|^{1-q} \leq R_q \eta^{1-q}. \quad (34)$$

If we set $\eta = \frac{\lambda_n}{c_1 k_1^2}$, by using (33) and (34) in (32) we have

$$\|\hat{f}_{\lambda_n} - f^*\|_2^2 \leq 13 R_q \left(\frac{\lambda_n}{c_1 k_1^2}\right)^{2-q} + \frac{2c_2 k_2^2 \log p}{c_1 k_1^2 n} R_q^2 \left(\frac{\lambda_n}{c_1 k_1^2}\right)^{2-2q}. \quad (35)$$

By Assumption 2.3, $T < M_1 \left(\frac{n}{R_q \log p}\right)^{\frac{2}{q}} \log p \log n$ and $\lambda_n = 2\sqrt{\frac{512M \log n \log p}{T}}$ we have

$$\frac{2c_2 k_2^2 \log p}{c_1 k_1^2 n} R_q^2 \left(\frac{\lambda_n}{c_1 k_1^2}\right)^{2-2q} \leq c_{M_1} R_q \left(\frac{\lambda_n}{c_1 k_1^2}\right)^{2-q},$$

where $c_{M_1} > 0$ is some constant depends on M_1 . Then from (35) with probability at least $1 - \frac{2}{p-1}$ there exists a $C_U > 0$ such that

$$\|\hat{f}_{\lambda_n} - f^*\|_2^2 \leq C_U R_q \left(\frac{\log n \log p}{T}\right)^{1-\frac{q}{2}}.$$

□

4.3 Proof for Theorem 3

The proof for Theorem 3 uses techniques developed in Raskutti et al. (2010) adapted from Gaussian to sub-Gaussian ensembles. The reason we adapt to sub-Gaussian ensembles is so that we construct a random ensemble that satisfies all the physical constraints. In the proof of Raskutti et al. (2010) the first step is to show the term $M(r, \Gamma) := \sup_{x \in V(r)} \{1 - \frac{\|\Gamma x\|_2}{\sqrt{n}}\}$ is sharply concentrated around its expectation with high probability when Γ is a matrix with gaussian random variables, we will use [Mendelson et al. (2007), Theorem 2.3] to show this is also true when Γ is a matrix with subgaussian random variables. Finally we use peeling techniques to complete the proof, which are used in Raskutti et al. (2010).

To begin we define the standard Gaussian width of a star-shaped set T . A set T is *start-shaped* if $cT \subset T$ for all $0 \leq c \leq 1$.

Definition 4.1 (Mendelson et al. (2007), Definition 2.1). Let $T \subset \mathbb{R}^p$ and let g_1, \dots, g_p be independent standard Gaussian random variables. Denote by $\ell_*(T) = \mathbb{E} \sup_{t \in T} |\sum_{i=1}^p g_i t_i|$, where $t = (t_i)_{i=1}^p \in \mathbb{R}^p$.

Now we state following result which is a restricted eigenvalue condition for random matrices with independent entries where each entry is an isotropic ψ_2 probability measure:

Theorem 4 (Mendelson et al. (2007), Theorem 2.3). *There exist absolute constants $c, \bar{c} > 0$ for which the following holds. Let $T \subset \mathbb{R}^p$ be a star-shaped set and put μ to be an isotropic ψ_2 probability measure with constant $\alpha \geq 1$. For $n \geq 1$ let X_1, \dots, X_n be independent, distributed according to μ and define $\Gamma = \sum_{i=1}^n \langle X_i, \cdot \rangle e_i$. If $0 < t < 1$, then with probability at least $1 - \exp(-\bar{c}f^2n/\alpha^4)$, for all $x \in T$ such that $\|x\|_2 \geq r_n^*(t/c\alpha^2)$, we have*

$$(1-t)\|x\|_2 \leq \frac{\|\Gamma x\|_2}{\sqrt{n}} \leq (1+t)\|x\|_2, \quad (36)$$

where

$$r_n^*(f) = r_n^*(f, T) := \inf\{\rho > 0 : \rho \geq \ell_*(T_\rho)/(f\sqrt{n})\}$$

and

$$T_\rho = \{x \in T; \|x\|_2 \leq \rho\}.$$

Next we want to prove the restricted eigenvalue condition for subgaussian random matrices by using this theorem.

Proof. We first note that it is sufficient to prove this theorem for $\|x\|_2 = 1$. In fact if $x = \mathbf{0} \in \mathbb{R}^p$ Theorem 3 holds trivially. Otherwise we can consider the re-scaled vector $\tilde{x} = x/\|x\|_2$ with $\|\tilde{x}\|_2 = 1$. It can be seen that if this theorem holds for the re-scaled vector \tilde{x} , it also holds for x .

Next we define the set $V(r) := \{x \in \mathbb{R}^p \mid \|x\|_2 = 1, \|x\|_1 \leq r\}$, for a fixed radius $r > 0$. It is possible that this set is empty for some choices of $r > 0$, but we only concern those choices for which it is non-empty. Define the random variable:

$$M(r, \Gamma) := \sup_{x \in V(r)} \left\{1 - \frac{\|\Gamma x\|_2}{\sqrt{n}}\right\}.$$

Our goal is to show that with probability no larger than $\exp(-c_\alpha n f(r)^2)$ that

$$M(r, \Gamma) = \sup_{x \in V(r)} \left\{1 - \frac{\|\Gamma x\|_2}{\sqrt{n}}\right\} \geq \frac{3f(r)}{2},$$

where $f(r) = \frac{1}{4} + 3c\alpha^2 r \sqrt{\frac{\log p}{n}}$, c_α and c are positive constants.

To see this, we choose $f = f(r) = \frac{1}{4} + 3c\alpha^2 r \sqrt{\frac{\log p}{n}}$ in Theorem 4 (this c here is defined in Theorem 4). We first bound $\ell_*(V(r))$ as follows:

$$\ell_*(V(r)) \leq r \mathbb{E} \max_{1 \leq i \leq p} |g_i| \leq 3r \sqrt{\log p}$$

by using known results on Gaussian maxima (Ledoux and Talagrand (1991), Equation (3.13)).

Then for all $x \in V(r)$ we have

$$\|x\|_2 = 1 \geq \frac{3c\alpha^2 r \sqrt{\frac{\log p}{n}}}{f(r)} \geq \frac{c\alpha^2 \ell_*(V(r))}{f(r) \sqrt{n}} = \frac{\ell_*(V(r))}{\frac{f(r)}{c\alpha^2} \sqrt{n}},$$

and by Theorem 4 with probability at least $1 - \exp(-\bar{c}f(r)^2 n / \alpha^4)$ we have for all $x \in V(r)$

$$1 - f(r) \leq \frac{\|\Gamma x\|_2}{\sqrt{n}}.$$

Hence with probability no larger than $\exp(-9\bar{c}nf(r)^2/4\alpha^4)$,

$$M(r, \Gamma) = \sup_{x \in V(r)} \left\{ 1 - \frac{\|\Gamma x\|_2}{\sqrt{n}} \right\} \geq \frac{3f(r)}{2}.$$

The remainder of the proof will mainly follow steps in Raskutti et al. (2010) where we use a peeling technique to extend our result to hold for x 's that have a random radius. We define the event

$$\Upsilon := \{\exists x \in \mathbb{R}^p \text{ s.t. } \|x\|_2 = 1 \text{ and } (1 - \|\Gamma x\|_2)/\sqrt{n} \geq 3f(\|x\|_1)\}.$$

To proof the main theorem, the next step is to show that there are positive constants c', c'' such that $P[\Upsilon] \leq c' \exp(-c''n)$. Now we follow the standard peeling technique (van de Geer (2000); Alexander (1985)) and we state the following lemma which is stated and proven in Raskutti et al. (2010).

Lemma 7 (Raskutti et al. (2010), Lemma 3). *Suppose that $d(v; X)$ is a random objective function with $v \in \mathbb{R}^p$ and X is some random vector. $h : \mathbb{R}^p \rightarrow \mathbb{R}_+$ is some non-negative and increasing constraint function. $g : \mathbb{R} \rightarrow \mathbb{R}_+$ is a non-negative and strictly increasing function. A is a non-empty set. Moreover we suppose $g(r) \geq u$ for all $r \geq 0$, and there exists some constant $c > 0$ such that for all $r > 0$, we have the tail bound*

$$\mathbb{P}\left(\sup_{v \in A, h(v) \leq r} d(v; X) \geq g(r)\right) \leq 2 \exp(-ca_n g^2(r)),$$

for some $a_n > 0$. Then we have

$$\mathbb{P}[\mathcal{E}] \leq \frac{2 \exp(-4ca_n u^2)}{1 - \exp(-4ca_n u^2)},$$

where

$$\mathcal{E} := \{\exists v \in A \text{ such that } d(v; X) \geq 2g(h(v))\}.$$

In order to use this lemma we choose the sequence $a_n = n$ and the set $A = \{x \in \mathbb{R}^p \mid \|x\|_2 = 1\}$, moreover we set

$$d(x, \Gamma) = 1 - \|\Gamma x\|_2 / \sqrt{n}, \quad h(x) = \|x\|_1, \quad \text{and} \quad g(r) = 3f(r)/2.$$

Since $g(r) = \frac{3f(r)}{2} = \frac{3}{8} + \frac{9}{2}c\alpha^2 r \sqrt{\frac{\log p}{n}} \geq \frac{3}{8}$ for all $r > 0$ and is strictly increasing, so that Lemma 7 is applicable with $u = \frac{3}{8}$. Thus by using Lemma 7 we have that $P[\Upsilon^c] \geq 1 - c' \exp(-c''n)$ for some numerical constants c' and c'' .

Then for all $x \in \mathbb{R}^p$ with $\|x\|_2 = 1$, conditioned on the event Υ^c we have

$$1 - \frac{\|\Gamma x\|_2}{\sqrt{n}} \leq 3f(\|x\|_1) = \frac{3}{4} + 9c\alpha^2 \|x\|_1 \sqrt{\frac{\log p}{n}},$$

then

$$\frac{\|\Gamma x\|_2}{\sqrt{n}} \geq \frac{1}{4} - 9c\alpha^2 \|x\|_1 \sqrt{\frac{\log p}{n}},$$

which completes the proof. \square

5 Appendix

5.1 Proof of Lemma 1

The proof for this lemma is mainly based on results in Vershynin (2010). First by using Lemma 5.5 in Vershynin (2010) we know that the definition of ψ_2 norm in Vershynin (2010) is equivalent with our definition up to some absolute constant. Then we use similar technique for the proof of Lemma 5.24 in Vershynin (2010). For every $x = (x_1, \dots, x_{p-1}) \in S^{p-2}$ we have

$$\|\langle X \bar{D}, x \rangle\|_{\psi_2}^2 = \|\langle X, x \bar{D}^\top \rangle\|_{\psi_2}^2 \leq C \sum_{i=1}^p (x \bar{D}^\top)_i^2 \|X_i\|_{\psi_2}^2 \leq C \max_{1 \leq i \leq p} \|X_i\|_{\psi_2}^2,$$

where the first inequality comes from Lemma 5.9 in Vershynin (2010) and we also used $\sum_{i=1}^p (x\bar{D}^\top)_i^2 = (x\bar{D}^\top)(x\bar{D}^\top)^\top = xx^\top = 1$ since $x \in S^{p-2}$. Since $X = (X_1, \dots, X_p)$ is distributed according to μ , we have shown that $\|\langle X\bar{D}, x \rangle\|_{\psi_2}$ is bounded by some absolute constant for every $x \in S^{p-2}$. It is also easy to see that $X\bar{D}$ is isotropic since X is isotropic and \bar{D} is orthonormal.

5.2 Proof of Lemma 5

To bound $\|\frac{2n}{T}(y - TAD\theta^*)^\top A\bar{D}\|_\infty$ we need the following two lemmas. Lemma 8 gives a concentration bound for poisson random variable and Lemma 9 gives a bound for $\sum_{i=1}^n (AD)_{ij}^2$, with $2 \leq j \leq p$.

Lemma 8. *Under Assumption 2.1 and 2.3, with probability at least $1 - \frac{2}{n}$,*

$$|\frac{n}{T}(y_i - T(AD\theta^*)_i)| = |\frac{n}{T}(y_i - T \sum_{j=1}^p A_{ij}f_j^*)| \leq \sqrt{\frac{32n \log n}{T}} \text{ for all } 1 \leq i \leq n.$$

The proof for Lemma 8 involves two-sided concentration bounds for Poisson random variables in combination with the union bound.

Lemma 9. *Under Assumption 2.2 we have for $2 \leq j \leq p$,*

$$\sum_{i=1}^n (AD)_{ij}^2 \leq \frac{(1 + \delta_{\tilde{K}})}{4n(a_u - a_\ell)^2}.$$

Proof. By the definition of \bar{D} and the construction of A

$$A\bar{D} = \left(\frac{\tilde{A} + \frac{a_u - 2a_\ell}{\sqrt{n}} \mathbb{1}_{n \times p}}{2(a_u - a_\ell)\sqrt{n}} \right) \bar{D} = \frac{\tilde{A}\bar{D}}{2(a_u - a_\ell)\sqrt{n}}.$$

Then for $2 \leq j \leq p$ we choose $u = e_j \in \mathbb{R}^p$ with j -th location to be 1 and all others to be 0, by Assumption 2.2,

$$\sum_{i=1}^n (AD)_{ij}^2 = \|A\bar{D}e'_j\|_2^2 = \frac{\|\tilde{A}\bar{D}e'_j\|_2^2}{4(a_u - a_\ell)^2 n} = \frac{\|\tilde{A}De_j\|_2^2}{4(a_u - a_\ell)^2 n} \leq \frac{(1 + \delta_{\tilde{K}})}{4n(a_u - a_\ell)^2},$$

where $e'_j \in \mathbb{R}^{p-1}$ is just e_j without the first element 0. □

Returning to the proof of Lemma 5, define $\bar{w}_i = \frac{n}{T}(y_i - T(AD\theta^*)_i)$ and

$$\begin{aligned} P(\|\frac{2n}{T}(y - TAD\theta^*)^\top A\bar{D}\|_\infty > t) &= P(\max_{2 \leq j \leq p} |\frac{2n}{T} \sum_{i=1}^n (y_i - T(AD\theta^*)_i)(AD)_{ij}| > t) \\ &= P(\max_{2 \leq j \leq p} |\sum_{i=1}^n \bar{w}_i(AD)_{ij}| > \frac{t}{2}). \end{aligned}$$

Since for $1 \leq i \leq n$, $\bar{w}_i(AD)_{ij} \in [-\sqrt{\frac{32n \log n}{T}}|(AD)_{ij}|, \sqrt{\frac{32n \log n}{T}}|(AD)_{ij}|]$ by Lemma 8, then by using Hoeffding's inequality and Lemma 9 we have

$$P(|\sum_{i=1}^n \bar{w}_i(AD)_{ij}| \geq \frac{t}{2}) \leq 2 \exp(-\frac{2(\frac{t}{2})^2}{\frac{128n \log n}{T} \sum_{i=1}^n (AD)_{ij}^2}) \leq 2 \exp(-\frac{t^2 T}{256M \log n}),$$

where $M = \frac{1+\delta_{\bar{K}}}{4(a_u - a_\ell)^2}$. By the union bound,

$$P(\max_{2 \leq j \leq p} |\sum_{i=1}^n \bar{w}_i(AD)_{ij}| > \frac{t}{2}) \leq 2 \exp(-\frac{t^2 T}{256M \log n} + \log(p-1)).$$

If we set $t = \sqrt{\frac{512M \log n \log(p-1)}{T}}$ then

$$P(\|\frac{2n}{T}(y - TAD\theta^*)^\top A\bar{D}\|_\infty > t) = \sqrt{\frac{512M \log n \log(p-1)}{T}} \leq \frac{2}{p-1}. \quad (37)$$

Thus with probability at least $1 - \frac{2}{p-1}$,

$$\|\frac{2n}{T}(y - TAD\theta^*)^\top A\bar{D}\|_\infty \leq \sqrt{\frac{512M \log n \log p}{T}}.$$

5.3 Proof of Lemma 6

Since $\hat{\Delta} \in \{\Delta \in \mathbb{R}^{p-1} \mid \|\Delta_{S_\eta^c}\|_1 \leq 3\|\Delta_{S_\eta}\|_1 + 4\|\bar{\theta}_{S_\eta^c}^*\|_1\}$, we have

$$\begin{aligned} \|\hat{\Delta}\|_1 &\leq 4\|\hat{\Delta}_{S_\eta}\|_1 + 4\|\bar{\theta}_{S_\eta}^*\|_1 \\ &\leq 4\sqrt{|S_\eta|}\|\hat{\Delta}\|_2 + 4\|\bar{\theta}_{S_\eta}^*\|_1 \\ &\leq 4\sqrt{R_q}\eta^{-q/2}\|\hat{\Delta}\|_2 + 4\|\bar{\theta}_{S_\eta}^*\|_1, \end{aligned}$$

where the third inequality follows from (33). Then by Assumption 2.4,

$$\begin{aligned} \sqrt{n}\|A\bar{D}\hat{\Delta}\|_2 &\geq k_1\|\hat{\Delta}\|_2 - 4k_2\sqrt{\frac{\log p}{n}}(\sqrt{R_q}\eta^{-q/2}\|\hat{\Delta}\|_2 + \|\bar{\theta}_{S_\eta}^*\|_1) \\ &= \|\hat{\Delta}\|_2(k_1 - 4k_2\sqrt{\frac{R_q \log p}{n}}\eta^{-q/2}) - 4k_2\sqrt{\frac{\log p}{n}}\|\bar{\theta}_{S_\eta}^*\|_1. \end{aligned}$$

By choosing $\eta = \sqrt{\frac{\log n \log p}{T}}$,

$$\sqrt{n}\|A\bar{D}\hat{\Delta}\|_2 \geq (k_1 - 4k_2)\sqrt{\frac{R_q \log p}{n}}\left(\frac{T}{\log p \log n}\right)^{q/4}\|\hat{\Delta}\|_2 - 4k_2\sqrt{\frac{\log p}{n}}\|\bar{\theta}_{S_\eta}^*\|_1.$$

Recall that Assumption 2.3 states that $T < M_1\left(\frac{n}{R_q \log p}\right)^{\frac{2}{q}} \log p \log n$ which implies

$$\sqrt{n}\|A\bar{D}\hat{\Delta}\|_2 \geq c'k_1\|\hat{\Delta}\|_2 - c''k_2\sqrt{\frac{\log p}{n}}\|\bar{\theta}_{S_\eta}^*\|_1,$$

where $c', c'' > 0$ are some constants. We can then find constants $c_1, c_2 > 0$ such that

$$n\|A\bar{D}\hat{\Delta}\|_2^2 \geq c_1k_1^2\|\hat{\Delta}\|_2^2 - c_2k_2^2\frac{\log p}{n}\|\bar{\theta}_{S_\eta}^*\|_1^2.$$

5.4 Proof of Lemma 8

To prove Lemma 8 we need the following result which is a two-sided version of the one-sided concentration bound from Bobkov and Ledoux (1998):

Lemma 10. *If $W \sim \text{Poisson}(\lambda)$, for any $t > 0$ we will have*

$$P(|W - \lambda| > t) \leq 2 \exp\left(-\frac{t}{4} \log\left(1 + \frac{t}{2\lambda}\right)\right).$$

Proof. First assume $t > 0$ and $u > 0$ and we have

$$\begin{aligned} P(W - \lambda > t) &= P(u(W - \lambda) > ut) \\ &= P(e^{u(W-\lambda)} > e^{ut}) \\ &\leq \frac{E(e^{u(W-\lambda)})}{e^{ut}} \\ &= \frac{E(e^{uW})}{e^{u(t+\lambda)}} \\ &= e^{\lambda(e^u - 1 - u) - ut}. \end{aligned}$$

Let $f(u) = \lambda(e^u - 1 - u) - ut$, the minimizer for $f(u)$ is $u^* = \log(1 + \frac{t}{\lambda})$. Hence

$$P(W - \lambda > t) \leq e^{\lambda(e^{u^*} - 1 - u^*) - u^*t} = e^{t - \lambda \log(1 + \frac{t}{\lambda}) - t \log(1 + \frac{t}{\lambda})},$$

since

$$t - \lambda \log\left(1 + \frac{t}{\lambda}\right) - t \log\left(1 + \frac{t}{\lambda}\right) \leq -\frac{t}{4} \log\left(1 + \frac{t}{2\lambda}\right),$$

we have

$$P(W - \lambda > t) \leq \exp(-\frac{t}{4} \log(1 + \frac{t}{2\lambda})).$$

On the other hand we assume $v > 0$ and $0 < t \leq \lambda$ which implies

$$\begin{aligned} P(W - \lambda < -t) &= P(\lambda - W > t) \\ &= P(v(\lambda - W) > vt) \\ &= P(e^{v(\lambda - W)} > e^{vt}) \\ &\leq \frac{Ee^{v(\lambda - W)}}{e^{vt}} \\ &= e^{\lambda(e^{-v} - 1 + v) - vt} \end{aligned}$$

Let $g(v) = \lambda(e^{-v} - 1 + v) - vt$, the minimizer for $g(v)$ is $v^* = -\log(1 - \frac{t}{\lambda})$. Thus we have

$$P(W - \lambda < -t) \leq e^{\lambda(e^{-v^*} - 1 + v^*) - v^*t} = e^{(t - \lambda) \log(1 - \frac{t}{\lambda}) - t},$$

since when $0 < t \leq \lambda$

$$(t - \lambda) \log(1 - \frac{t}{\lambda}) - t \leq -\frac{t}{4} \log(1 + \frac{t}{2\lambda}),$$

we have

$$P(W - \lambda < -t) \leq \exp(-\frac{t}{4} \log(1 + \frac{t}{2\lambda})).$$

Note that if $t > \lambda$ the inequality above is always true. Combining these two one-sided results together we get the complete proof for Lemma 10. \square

Returning to the proof of Lemma 8, we have

$$\begin{aligned} P(|\frac{n}{T}(y_i - T(AD\theta^*)_i)| > \sqrt{\frac{32n \log n}{T}}) &= P(|y_i - T \sum_{j=1}^p A_{ij} f_j^*| > \sqrt{\frac{32T \log n}{n}}) \\ &\leq 2 \exp(-\frac{1}{4} \sqrt{\frac{32T \log n}{n}} \log(1 + \frac{\sqrt{\frac{32T \log n}{n}}}{\frac{2T}{n}})) \\ &= 2 \exp(-\frac{1}{4} \sqrt{\frac{32T \log n}{n}} \log(1 + \frac{1}{2} \sqrt{\frac{32n \log n}{T}})). \end{aligned}$$

The second inequality follows from Lemma 10 and $\frac{T}{2n} \leq T(Af^*)_i \leq \frac{T}{n}$ from Lemma 4. By using Assumption 2.3 and fact that $\log(1 + x) > \frac{x}{2}$ for $x \in (0, 1)$,

$$P(|\frac{n}{T}(y_i - T(AD\theta^*)_i)| > \sqrt{\frac{32n \log n}{T}}) \leq 2 \exp(-\frac{32 \log n}{16}) = 2 \exp(-2 \log n).$$

Then

$$P\left(\max_{1 \leq i \leq n} \left| \frac{n}{T} (y_i - T(AD\theta^*)_i) \right| > \sqrt{\frac{32n \log n}{T}}\right) \leq 2n \cdot \exp(-2 \log n) = \frac{2}{n}.$$

This completes the proof for Lemma 8.

References

- K. S. Alexander. Rates of growth for weighted empirical processes. In *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, pages 475–493, Berkeley, 1985. UC Press.
- R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, 28(3):253–263, 2008. ISSN 0176-4276. doi: 10.1007/s00365-007-9003-x. URL <http://dx.doi.org/10.1007/s00365-007-9003-x>.
- P. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. Submitted to *Annals of Statistics*, 2008.
- P. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37(4):1705–1732, 2009.
- J. Bobin, J.-L. Starck, J. Fadili, and Y. Moudden. Sparsity and morphological diversity in blind source separation. *IEEE Transactions on Signal Processing*, 16(11):2662–2674, 2007a.
- J. Bobin, J.-L. Starck, J. Fadili, Y. Moudden, and D. L. Donoho. Morphological component analysis: An adaptive thresholding strategy. *IEEE Transactions on Image Processing*, 16(11):2675–2681, 2007b.
- S. G. Bobkov and M. Ledoux. On modified logarithmic sobolev inequalities for bernoulli and poisson measures. *Journal of Functional Analysis*, 156(2):347–365, 1998.
- M. F. Duarte, M. A. Davenport, D. Takhar, J. N. Laska, T. Sun, K. F. Kelly, and R. G. Baraniuk. Single pixel imaging via compressive sampling. *IEEE Sig. Proc. Mag.*, 25(2): 83–91, 2008.

- C. Estan and G. Varghese. New directions in traffic measurement and accounting: focusing on the elephants, ignoring the mice. *ACM Trans. Computer Sys.*, 21(3):270–313, 2003.
- T. S. Han and S. Verdú. Generalizing the fano inequality. *IEEE Transactions on Information Theory*, 40(4):1247–1251, 1994.
- Z. T. Harmany, J. Mueller, Q. Brown, N. Ramanujam, and R. Willett. Tissue quantification in photon-limited microendoscopy. In *Proc. SPIE Optics and Photonics*, 2011.
- W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.
- I. A. Ibragimov and R. Z. Has'minskii. *Statistical Estimation: Asymptotic Theory*. Springer-Verlag, New York, 1981.
- X. Jiang, G. Raskutti, and R. Willett. Minimax optimal rates for poisson inverse problems with physical constraints. *arXiv preprint arXiv:1403.6532*, 2014.
- X. Jiang, P. Reynaud-Bouret, V. Rivoirard, L. Sansonnet, and R. Willett. A data-dependent weighted lasso under poisson noise. *arXiv preprint arXiv:1509.08892*, 2015.
- T. Kühn. A lower estimate for entropy numbers. *Journal of Approximation Theory*, 110:120–124, 2001.
- M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer-Verlag, New York, NY, 1991.
- Y. Lu, A. Montanari, B. Prabhakar, S. Dharmapurikar, and A. Kabbani. Counter Braids: a novel counter architecture for per-flow measurement. In *Proc. ACM SIGMETRICS*, 2008.
- S. Mendelson, A. Pajor, and N. Tomczak-Jaegermann. Reconstruction and subgaussian operators in asymptotic geometric analysis. *Geometric and Functional Analysis*, 17(4):1248–1282, 2007.
- S. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of m -estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

- M. Raginsky, S. Jafarpour, R. Willett, and R. Calderbank. Fishing in Poisson streams: focusing on the whales, ignoring the minnows. In *Proc. Forty-Fourth Conference on Information Sciences and Systems*, 2010a. arXiv:1003.2836.
- M. Raginsky, R. Willett, Z. Harmany, and R. Marcia. Compressed sensing performance bounds under Poisson noise. *IEEE Transactions on Signal Processing*, 58(8):3990–4002, 2010b. arXiv:0910.5146.
- M. Raginsky, S. Jafarpour, Z. Harmany, R. Marcia, R. Willett, and R. Calderbank. Performance bounds for expander-based compressed sensing in Poisson noise. *IEEE Transactions on Signal Processing*, 59(9), 2011. arXiv:1007.2377.
- G. Raskutti, M. J. Wainwright, and B. Yu. Restricted eigenvalue conditions for correlated Gaussian designs. *Journal of Machine Learning Research*, 11:2241–2259, 2010.
- G. Raskutti, M. J. Wainwright, and B. Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q -balls. *IEEE Transactions of Information Theory*, 57(10):6976–6994, 2011.
- L. Sansonnet. Wavelet thresholding estimation in a Poissonian interactions model with application to genomic data. *Scandinavian Journal of Statistics*, 2013. doi: 10.1111/sjos.12009.
- V. Studer, J. Bobin, M. Chahid, H. S. Mousavi, E. Candes, and M. Dahan. Compressive fluorescence microscopy for biological and hyperspectral imaging. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, 109(26), 2012.
- S. van de Geer. *Empirical Processes in M-Estimation*. Cambridge University Press, 2000.
- S. van de Geer and P. Bühlmann. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.
- R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.
- R. M. Willett and M. Raginsky. Poisson compressed sensing. *Defense Applications of Signal Processing*, 2011.

- Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, 27(5):1564–1599, 1999.
- S. Zhou. Restricted eigenvalue conditions on subgaussian random matrices. *arXiv preprint arXiv:0912.4045*, 2009.